

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-203201
 (43)Date of publication of application : 30.07.1999

(51)Int.CI. G06F 12/08
 G06F 12/08

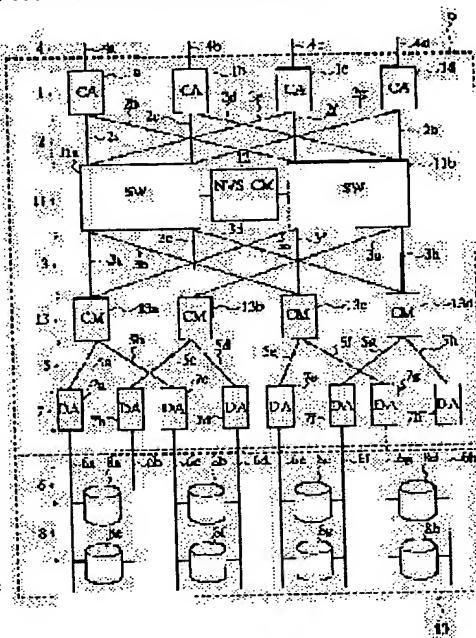
(21)Application number : 10-002400 (71)Applicant : HITACHI LTD
 (22)Date of filing : 08.01.1998 (72)Inventor : MORI KENJI

(54) ARRANGING METHOD OF CACHE MEMORY AND DATA STORAGE SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To dissolve bottleneck of an access to a cache memory caused by an increase of the number of paths or storage devices of a host device.

SOLUTION: A non-volatile cache memory (12) where write-in data from the host device are stored and a volatile cache memory 143 where data read from a disk drive 8 are temporarily stored are separately provided between plural channel I/F control circuits 1 that control plural channels I/F 4 on the side of the host device and plural disk control circuits 7 that control plural disk drives 8 in a disk drive unit 10. Moreover, the non-volatile cache memory (12) is commonly and concentratedly provided in plural data transmission paths, the volatile cache memory 13 is distributed every several data transmission paths and is arranged, and set of each capacity or throughput of the non-volatile cache memory (12) and plural volatile cache memories 13 is individually enabled.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

* NOTICES *

Japan Patent Office is not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

CLAIMS

[Claim(s)]

[Claim 1] It is the configuration method of the cache memory characterized by to be arranged between high order equipment and the storage with which the information delivered and received between the aforementioned high order equipment is stored, to be the configuration method of the cache memory which holds the aforementioned information temporarily, to constitute the aforementioned cache memory from a non-volatilized cache memory and a volatilization cache memory, to arrange the aforementioned non-volatilized cache memory intensively, and to distribute and arrange the aforementioned volatilization cache memory.

[Claim 2] It is arranged between high order equipment and the storage with which the information delivered and received between the aforementioned high order equipment is stored. It is the configuration method of the cache memory which holds the aforementioned information temporarily, and the aforementioned cache memory is constituted from a non-volatilized cache memory and a volatilization cache memory. The capacity of the aforementioned non-volatilized cache memory, The throughput of the 1st method of setting up so that the capacity of the aforementioned volatilization cache memory may differ, and the aforementioned non-volatilized cache memory, The configuration method of the cache memory characterized by using one [at least] method of 2nd method ** set up so that the throughputs of the aforementioned volatilization cache memory may differ.

[Claim 3] The data-storage system containing the cache memory in which it is arranged between the memory control unit which controls transfer of the aforementioned information between the storage with which the information delivered and received between the high order equipment characterized by to provide the following is stored, and the aforementioned storage and the aforementioned high order equipment, and the aforementioned high order equipment and the aforementioned storage, and the aforementioned information is stored temporarily. It is the capacity of the 1st composition and the aforementioned non-volatilized cache memory which distributes the aforementioned volatilization cache memory and is arranged by the aforementioned cache memory's consisting of a non-volatilized cache memory and a volatilization cache memory, and arranging the aforementioned non-volatilized cache memory intensively. The throughput of the 2nd composition from which the capacity of the aforementioned volatilization cache memory differs, and the aforementioned non-volatilized cache memory. At least one composition of 3rd composition ** from which the throughput of the aforementioned volatilization cache memory differs.

[Translation done.]

* NOTICES *

Japan Patent Office is not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. *** shows the word which can not be translated.
3. In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[The technical field to which invention belongs] Especially this invention is applied to the data-storage system by which informational transfer is performed in parallel in two or more information transfer paths established between high order equipment and storage about the arrangement technology and data-storage technology of a cache memory, and relates to effective technology.

[Description of the Prior Art] Drawing 13 is the conceptual diagram showing an example of the composition of a disk subsystem which consists of the conventional disk controller 109 and a subordinate's disk drive unit 110. The composition which arranges the cache memory accessed in the conventional disk controller 109 through the cache memory access path 106 between the channel-control circuit 101 which performs input/output control by the side of a host, and the disk control circuit 107 which performs input/output control by the side of a disk drive is common. At this time, the cache memory was used as a simple double system with the non-volatilized cache memory 102 with which the data loss resulting from a power failure was equipped, and the volatilization cache memory 103 as an object for the compensation at the time of the obstacle of a cache memory. In this case, when the path 104 to the opposite host of equipment and the number of disk drives 108 increase, in order for access to occur in all data accesses in two non-volatilized cache memories 102 and the volatilization cache memory 103 and to carry out access concentration at a cache memory, it becomes a bottleneck in the performance of a disk controller 109.

[0003] However, in the conventional disk controller 109, there are few the paths 104 and disk drives 108 to an opposite host like drawing 13 as four channels and 8 drive path grade, and since the performance of the grade which is such a method is securable, such a method has been taken.

[0004] [Problem(s) to be Solved by the Invention] The number of the disk drives which control will increase by the increase in the capacity per disk controller, and advanced features from now on, and it will be thought that the number of paths to the host of equipment increases.

[0005] When such a control unit is constituted from a conventional method, access to a cache memory concentrates and a cache memory serves as a bottleneck of a control unit. In order to cancel this bottleneck, how to distribute a cache memory can be considered. It is necessary to prepare the cache memory of the double system which contains a non-volatilized cache memory in preparation for an obstacle, respectively like the former, and at this time, if it is such and a distributed cache is constituted, it will become expensive also in price difficult in mounting.

[0006] The purpose of this invention is to offer the arrangement technology and data-storage technology of the cache memory which can raise the throughput of the data transfer which went via the cache memory.

[0007] Other purposes of this invention are to offer the arrangement technology and data-storage technology of the cache memory which can realize the cost reduction in the cache

memory of composition of that a non-volatilized cache memory and a volatilization cache memory are intermingled.

[0008] Other purposes of this invention are to offer the arrangement technology and data-storage technology of the cache memory which can realize improvement in the mounting efficiency in the cache memory of composition of that a non-volatilized cache memory and a volatilization cache memory are intermingled.

[0009]

[Means for Solving the Problem] In this invention, the non-volatilized cache memory for guaranteeing the data at the time of a light is arranged to one place. Moreover, in order to gather the efficiency of a lead, a volatilization cache memory is distributed so that two or more accesses can be performed. This volatilization cache memory distributed is used only at the time of a lead, and since it can lead from a disk even if data should volatilize by powering off etc., it can consist of volatile memory media.

[0010] Moreover, in memory control units, such as a disk controller, the ratio of read/write is not symmetrical and is used in many cases in about one light to lead 4. If this property is used, memory space will not influence to a performance, even if a non-volatilized cache memory does not enlarge the volatilization cache memory for a lead with the ratio of read/write. Moreover, a non-volatilized cache memory can also make a throughput smaller than the volatilization cache memory for a lead. However, since the capacity of the volatilization cache memory for a lead to non-distribute influences to a performance sensitively, it needs to make capacity larger than a non-volatilized cache memory. Thus, a cache memory can be made into the optimal amount by setting it as the value according to the performance of which each capacity is required in two groups of distribution and concentration of arrangement of a cache memory, and highly efficientization can be attained by the low price in favor also of a mounting price target by this.

[0011] From now on, the number of the disk drives which control increases by the increase in the capacity per disk controller, and advanced features, and it will be thought that the path to the host of equipment increases.

[0012] In this case, in this invention, a non-volatilized cache memory is intensively arranged to one place to two or more paths, and a volatilization cache memory is distributed for every one or some of paths. By distributing these volatilization cache memory, the transfer for a cache memory is attained simultaneously, and a high throughput can be brought about, without concentrating on one cache memory at the time of the correspondence to many channels and the formation of many drives. Moreover, since the cache memory to distribute is good for volatility, it is advantageous also in mounting, and it can consist of low prices. By making into a unit size of the volatilization cache memory which has an extension unit furthermore distributed, it becomes possible to take free throughput composition. It becomes possible to attain highly efficientization by the low price in the system of for example, two or more path composition by considering as the composition of such a cache memory.

[0013] [Embedments of the Invention] Hereafter, the gestalt of operation of this invention is explained in detail, referring to a drawing.

[0014] Drawing 1 is the conceptual diagram showing an example of the composition of the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced. The gestalt of this operation takes and explains the case where it applies to a disk subsystem to an example of a data-storage subsystem.

[0015] The disk subsystem of the gestalt of this operation illustrated by drawing 1 is large, and consists of a disk control unit 9 and two units of disk drive unit 10*, the channel I/F control circuit 1 (1a-1d) of plurality [control unit / disk / 9], two or more data path switches 11 (11a-11b), the non-volatilized cache memory 12, two or more volatilization cache memories 13 (13a-13d), and two or more disk control circuits 7 (7a-7h) — it is constituted more and the disk drive unit 10 consists of two or more disk drives 8 (8a-8h)

[0016] In drawing 1, it connects with host equipments which are not illustrated through two or more channel I/F4 (4a-4d), such as a channel unit and a central processing unit (CPU), individually, and two or more channel I/F control circuits [1a-1d] each performs protocol

control of channel I/F4, data conversion, and data transfer.

[0017] Two or more data path switches 11 achieve the duty which changes the path of the data flow between two or more channel I/F control circuits 1 connected through two or more path 2 (2a-2h), and two or more paths 3 (3a-3h), and the non-volatilized cache memory 12 and the volatilization cache memory 13.

[0018] Two or more volatilization cache memories 13 are connected to two or more disk control circuits 7 through two or more paths 5 (5a-5h), and two or more disk control circuits 7 are connected to the disk drive unit 10 through two or more paths 6 (6a-6h).

[0019] Since the data at the time of the light from a host side are stored, the non-volatilized cache memory 12 is used. This is for performing control which is the purpose of improvement in the speed of light processing, and returns completion before writing in a disk drive 8, when the writing to the non-volatilized cache memory 12 is completed to the light command from a host side, and before it writes data in a disk drive 8, it should be used for the purpose which backs up the data at the time of equipment stopping by power failure by this non-volatilized cache memory 12. Moreover, when writing in this non-volatilized cache memory 12, a path is set up with the data path switch 11 so that it may write in the volatilization cache memory 13 as simultaneous as possible.

[0020] When the volatilization cache memory 13 mainly works as a cache memory of lead data and leads data from a disk drive 8, it is used for the purpose which attains improvement in the speed by leading from the volatilization cache memory 13 rather than storing data in this volatilization cache memory 13 and leading data from a disk drive 8 to the lead over the same data as 2nd henceforth. For this reason, since data are saved at the disk drive even when stored in an volatile memory medium.

[0021] It connects with the disk drive 8 through the path 6, and the disk control circuit 7 is used for controlling a disk drive 8. When it has two paths for disk drive control and this uses a shift, path also at the time of emergency obstacle generating, access to a disk drive 8 is possible for one disk control circuit 7. Moreover, the function of improvement in the speed is also usually achieved by using a two pass effectively at the time.

[0022] A disk drive 8 is used for storing data, and the access path to a disk drive 8 has two paths for improvement in the speed and a raise in reliance.

[0023] That is, two or more disk drives 8 which constitute the disk drive unit 10 in the case of the gestalt of this operation are systematized by some groups (four as [The gestalt of this operation] an example), and each sequence is connected to the disk control unit 9 in the multiplex path respectively through two or more paths 6a-6b, Paths 6c-6d, Paths 6e-6f, and Paths 6g-6h. Moreover, the data transfer in two or more paths [6a-6h] each, is independently controlled in two or more disk control circuits [7a-7h] each.

[0024] Hereafter, the configuration method of the cache memory of the gestalt of this operation and an example of an operation of a disk subsystem are explained.

[0025] Drawing 2 is the conceptual diagram having shown an example of the data flow at the time of execution of the lead command from CPU. This drawing 2 explains the case where a data lead command goes into channel I/F control circuit 1a from CPU through channel I/F4a, for example.

[0026] First, the command from CPU is recognized by channel I/F control circuit 1a. It turns out that it is a lead command as a result. Then, it distinguishes of which disk drive 8 it is data changes in data path switch 11a, corresponding to a distinction result, and let channel I/F control circuit 1a and volatilization cache memory 13a be integrated states.

[0027] Next, it judges whether data are on a cache memory. This judgment is performed by channel I/F control circuit 1a. Consequently, when it judges with object data being on a cache memory (cache hit), it reads by transmitting to channel control circuit 1a data-path switch 11a channel I/F control circuit 1a. Through data from volatilization cache memory 13a like data flow 16, channel I/F control circuit 1a transmits data to a host (CPU) through channel I/F4a, and a lead command is completed.

[0028] When there are no data for access on volatilization cache memory 13a (cache mistake), disk drive 8 is led via disk control circuit 7a like data flow 17, and data are transmitted and read

through data path switch 11a. Under the present circumstances, the light of the data is simultaneously carried out to volatilization cache memory 13a. This becomes a cache hit to the data lead of the 2nd henceforth, and improvement in the speed of access can be attained.

[0029] Drawing 3 is the conceptual diagram having shown an example of the data flow at the time of execution of the light command from CPU. The case where a data light command goes into channel I/F control circuit 1a from CPU through channel I/F4a is explained.

[0030] The command from CPU is first recognized by channel I/F control circuit 1a. It turns out that it is a light command as a result. A light object distinguishes which disk drive 8 after this, and data path switch 11a is changed according to a distinction result, for example, let channel I/F control circuit 1a, the non-volatilized cache memory 12, and volatilization cache memory 13a be integrated states.

[0031] The data from channel I/F control circuit 1a are simultaneously transmitted to volatilization cache memory 13a and the non-volatilized cache memory 12 like data flow 21. Thereby, they are transmitted to distributed cache memory 13a while the light of the light data is carried out on the non-volatilized cache memory 12. Data are written in the disk drive 8a concerned by transmitting light data also to disk control circuit 7a through path 5a next, and being further transmitted to disk drive 8a of the purpose through path 6a.

[0032] Next, operation when a power failure occurs is explained. Channel I/F control circuit 1a receives data, and when a transfer is completed to the non-volatilized cache memory 13a while the light of the light data is data flow 22, channel I/F control circuit 1a returns command completion to CPU. For this reason, it will be judged in CPU that data writing was completed. However, at this time, data are in the disk control unit 9, and it is in the state where it is not written in disk drive 8a yet. If a power failure occurs at this time, the data on volatilization cache memory 13a will disappear among on the non-volatilized cache memory 12 which the transfer completed, and volatilization cache memory 13a. Then, when a power supply is restored, it checks whether non-written in data exist on the non-volatilized cache memory 12, and the writing to disk drive 8a of the data in the disk control unit 9 is resumed. At this time, the data from the non-volatilized cache memory 12 are transmitted to disk drive 8a of the purpose through path 3a from data path switch 11a, 12 are transmitted to disk control circuit 7a of data path 5a, disk control circuit 7a, and path 6a like data flow 23.

[0033] Drawing 4 is the conceptual diagram showing an example of the data flow at the time of two or more read/write generating.

[0034] This is drawing showing an example of processing when channel I/F control circuit 1a has the demand of a lead command through channel I/F 4b, 4c, and 4d at the channel I/F control circuits 1b, 1c, and 1d of others [demand / of a light command] through channel I/F4a.

[0035] First, operation of channel I/F control circuit 1a is explained. In order that channel I/F control circuit 1a may execute a light command, it controls data path switch 11a, and chooses the path to the non-volatilized cache memory 12 and volatilization cache memory 13a. And like data flow 34, the light data from CPU are transmitted to the non-volatilized cache memory 12 via path 2a and data path switch 11a, are transmitted to disk control circuit 7a through path 2a, data path switch 11a, path 3a, volatilization cache memory 13a, and path 5a, and are simultaneously written further in disk drive 8a via path 6a.

[0036] Next, operation of the channel I/F control circuits 1b and 1c is explained. In order to execute a lead command by channel I/F control circuit 1b, data path switch 11a chooses path switch 11b chooses volatilization cache memory 13b. At this time, data path switch 11a can form two kinds of paths simultaneously according to two demands, channel I/F control circuit 1a and channel I/F control circuit 1b. And when the data for a lead exist in volatilization cache memory 13b (cache hit), data transfer is performed via path 3c and path 2c from volatilization cache memory 13b like data flow 35 to channel I/F control circuit 1b.

[0037] Moreover, in channel I/F control circuit 1c, in order to execute a lead command, data path switch 11b chooses volatilization cache memory 13c. And when the data for a lead exist in path switch 11b chooses volatilization cache memory 13c (cache hit), data transfer is performed like data flow 36 to channel I/F control circuit 1c via path 3f from volatilization cache memory 13c, and path 2f.

[0038] Finally operation of 1d of channel I/F control circuits is explained. In order to execute a lead command by 1d of channel I/F control circuits, data path switch 11b chooses volatilization

cache memory 13d. When object data do not exist in this volatilization cache memory 13d (cache mistake), a lead demand goes into 7h of disk control circuits via path 5h like data flow 37, and the lead data read from disk drive 8d via path 6h are transmitted to 1d of channel 1/F control circuits via path 3h and path 2h, while a light is carried out to volatilization cache memory 13d. Thus, it becomes possible by distributing the volatilization cache memory 13 like 13a-13d for every path to perform simultaneously data transfer to two or more paths.

[0043] Drawing 5 is the conceptual diagram showing an example of data flow when two or more read/write concentrates on a specific disk drive (for example, disk drive 8d).

[0040] This is drawing when the channel 1/F control circuits 1a and 1b have the demand of a lead command at a light command and the channel 1/F control circuits 1c and 1d.

[0041] Operation of channel 1/F control circuit 1a is explained first. Like data flow 43, channel 1/F control circuit 1a chooses the path to the non-volatile cache memory 12 and volatilization cache memory 13c for data path switch 11a in order to execute a light command. However, when 7g of disk control circuits is used by the lead command etc. at this time, after it performs only the transfer to the non-volatile cache memory 12 previously and 7g of disk control circuits is vacant after that, like the time of a power failure, from the non-volatile cache memory 12, data are transmitted to 7g of disk control circuits via volatilization cache memory 13c, and it writes in disk drive 8d. Since 7g of disk control circuits is vacant in the case of the example of drawing 5, it transmits simultaneously.

[0042] The end report of a command to a host is performed, when it writes in the non-volatile cache memory 12, before writing was completed to disk drive 8d. It writes in the non-volatile cache memory 12 certainly fundamentally, and then, when the 7g [of disk control circuits] path is vacant, it considers as the processing which performs the transfer to 7g of disk control circuits simultaneously.

[0043] After it writes in the non-volatile cache memory 12 temporarily like data flow 44 similarly about operation of channel 1/F control circuit 1b and 7h of disk control circuits is vacant after that, from the non-volatile cache memory 12, data are transmitted to 7h of disk control circuits via volatilization cache memory 13d, and it writes in disk drive 8d. In this example, although, as for the writing to the non-volatile cache memory 12, a light demand [control circuits / channel 1/F / 1a and 1b / two] competes, this is performed in order.

[0044] Next, operation of channel 1/F control circuit 1c is explained. In order to execute a lead command by channel 1/F control circuit 1c, data path switch 11b chooses volatilization cache memory 13d. And when the data for a lead exist in volatilization cache memory 13d (cache hit), data transfer is performed like data flow 45 to channel 1/F control circuit 1c via volatilization cache memory 13d to path 3h, and path 2h.

[0045] When the target data do not exist in volatilization cache memory 13d (cache mistake), data are read from disk drive 8d via path 6h, 7h of disk control circuits, path 5h, volatilization cache memory 13d, and path 3h.

[0046] It checks whether the data for a lead are in volatilization cache memory 13d first like data flow 46 similarly also about 1d also of channel 1/F control circuits, and when it transmits in a certain case (cache hit) and there is nothing to it from volatilization cache memory 13d (cache mistake), data are read from disk drive 8d. Although a lead command competes at this time, this is performed in order.

[0047] In the Prior art in such a case, the competition of access to a cache memory would occur, and waiting of access [the lead over a single disk drive and] to the disk drive which also distributed the light will increase, and they caused degradation.

[0048] However, with the form of this operation, in access to the dispersed disk drive, competition is not generated but high performance is obtained. Moreover, since there are few cases which access to one disk drive 8d concentrates like drawing 5 at transaction processing which much short accesses generate, it can be said that there is little concentration of access to the specific volatilization cache memory at the time of a lead. Moreover, although there is a non-volatile cache memory 12 as memory which carries out a centralized control and access concentrates at the time of a light, generally the ratio of a light does not become so much problem compared with a lead for a quadrant grade. Therefore, by this, there is also little

capacity of the non-volatile cache memory 12 compared with total of other volatilization cache memories 13a-13d, it ends, and it becomes possible to realize a highly efficient disk subsystem by the low price.

[0049] Explanation about the size of a cache memory and a throughput is given here. The case of the composition of having constructed four and having had which is 13a, 13b, 13c, and 13d considering one and the volatilization cache memory 13 as an example about the non-volatile cache memory 12 like drawing 5 is considered. The ratio to a light and a lead is set to 1:3. be fastidious -- it is alike, and the size of the non-volatile cache memory 12 and a throughput will live in a twist by the quadrant grade of the equivalent size of two or more volatilization cache memories 13 (13a-13d) used for a lead, and a throughput, and are convenient for it mounting-- wise and in price

[0050] Next, the size of the volatilization cache memory 13 which consists of two or more volatilization cache memories 13a-13d, and a throughput are considered. When each Carver range of two or more volatilization cache memories 13a-13d considers as a disk drive path (Paths 5a and 5b, Paths 5c and 5d, Paths 5e and 5f, Paths 5g and 5h) at this time, the value is decided by the degree of concentration to this disk drive path. It consists of 8 Paths [5a-5h] sets like drawing 5, and although the equivalent throughput itself is sufficient when each part is assumed, the effect of a cache memory can be pulled out by having that much mostly. This will determine a value probable.

[0051] By the way, with the composition which writes in a host side and answers completion when the non-volatile cache memory 12 and the volatilization cache memory 13 (13a-13d) are distributed and arranged and the data writing to the non-volatile cache memory 12 was completed like the gestalt of this operation, the write-in data to the non-volatile cache memory 12 are not necessarily immediately reflected in the volatilization cache memory 13 or the disk drive 8. For this reason, when a lead demand occurs between being un-reflected, the operation the newest data distinguish [operation] in any it shall exist, between the non-volatile cache memory 12, the volatilization cache memory 13, and a disk drive 8 is needed.

[0052] With the gestalt of this operation, such distinction operation is performed as an example using control information which is illustrated by drawing 6.

[0053] Namely, in the non-volatile cache memory 12, the NVS management flag 50 (VN) is formed, for example for every entry of an access unit. In the case of the gestalt of this operation, it is VN. When it is "0", the write-in data of the entry concerned have not been reflected in the volatilization cache memory 13, and it is reflection settled at the time of "1".

[0054] Moreover, in the volatilization cache memory 13, CM management flag 51 (V, A) is formed, for example for every entry of an access unit. When V is "0" in the case of the gestalt of this operation, non-reflected data exist in the non-volatile cache memory 12 to the data of the entry concerned, and when V is "1", it does not exist. Moreover, when A is "0", the write-in data of the entry concerned are in the state where it is not reflected in a disk drive 8 top, and it is reflection settled when A is "1".

[0055] In addition, in CM management flag 51, although storing data have disappeared immediately after powering on, in this state, both V and A of all entries are in the state of "0", it is judged with a cache mistake and the data lead from a disk drive 8 is performed by this state. And V and A change with the write-in operation to the disk drive 8 concerned of the data which are not reflected in the disk drive 8 which exists in the non-volatile cache memory 12, and storing operations of the data read from the disk drive 8 like the after-mentioned.

[0056] And VN of the NVS management flag 50 after writing the write-in data which come from a host side (channel 1/F control circuit 1) in the non-volatile cache memory 12 on the occasion of data writing, for example so that it may be illustrated by the flow chart of drawing 7 it sets to "0" (Step 20) and V of CM management flag 51 is further set to "0" (Step 202). Then, a host side is answered in light completion (Step 203). In addition, although access to the

volatilization cache memory 13 occurs at Step 202 for operation of V of CM management flag 51, since it is only operation of few flag bits unlike the usual data transfer, there are few overheads. [0057] For example, a 12 or less cache memory [non-volatile] write-in data transfer is performed like the example of above-mentioned drawing 5 by the procedure in which an arbitrary opportunity is sufficient, for example, the flow chart instantiation of drawing 8 is carried out. [0058] That is, it is VN of the NVS management flag 50 first. After ["0"] searching an entry from the non-volatile cache memory 12 (Step 301) and transmitting the data concerned to the volatilization cache memory 13, V of CM management flag 51 is set to "1" (Step 302). Furthermore, after writing in on a disk drive 8 from the volatilization cache memory 13 and transmitting data A of CM management flag 51 is set to "1" (Step 303). To the last, it is VN of the NVS management flag 50. It sets to "1" (Step 304). This the operation of a series of is an execute permission in an arbitrary opportunity.

[0059] Processing of the lead demand from the host side generated in arbitrary opportunities on the other hand is performed by carrying out like the flow chart illustrated by drawing 9 as an example.

[0060] That is, if a lead demand occurs, CM management flag 51 of the volatilization cache memory 13 which corresponds first is checked (Step 401), and in being A=1 and V=0, it will judge with the write-in data which are not reflected corresponding to the data by which the lead demand was carried out existing in the non-volatile cache memory 12, data will be read from the non-volatile cache memory 12, and it will transmit to a host (Step 404).

[0061] Moreover, in Step 401, when judged with it not being A=1 and V=0, it investigates further whether it is A=0 and V=1 or A=1, and V=1 (Step 402), and when this condition is satisfied, as a cache hit of the volatilization cache memory 13, the data in the volatilization cache memory 13 are read, and it transmits to a host side (Step 405).

[0062] Judging with a cache mistake, reading data from a disk drive 8, and writing in the volatilization cache memory 13, in agreeing on neither of the conditions, Step 401 nor Step 402, data are transmitted to a host side and A and V of CM management flag 51 are set to "1" (Step 403).

[0063] By a series of processings using such a NVS management flag 50 and CM management flag 51 Regardless of data write-in operation for a data write request it to have not performed on the level of disk drive 8 throat further [the 12 or less cache memory / non-volatile / volatilization cache memory 13 and] Generating of the obstacle of being an execute permission exactly, for example, the newest write-in data reading old non-reflected data accidentally, and transmitting the lead of the newest data to a host side is certainly avoidable to the lead demand from a host side.

[0064] Moreover, since the data accessed on the occasion of such management are at most several bits, the overhead resulting from operation of the NVS management flag 50 and CM management flag 51 hardly influences the throughput of read/write processing. [0065] As explained above, according to the configuration method and data-storage system of a cache memory of this operation, [of a gestalt], in the formation of a multi-channel path equipped with two or more channel I/F4 and paths 2, and the composition of the disk control unit 9 formed into many disk drives by increase of the quantity of the disk drive 8 in the disk drive unit 10 By distributing for some of every paths, the volatilization cache memory 13 It is high throughput-ization being attained, and arranging the non-volatile cache memory 12 intensively and managing it separately [the volatilization cache memory 13], further. Each size of the non-volatile cache memory 12 and the volatilization cache memory 13 can be set up subsystem, by the low price advantageous by the component side.

[0066] In addition, the composition illustrated by not only the method illustrated by drawing 1 but drawing 10 - drawing 12 as the distribution method of a non-volatile cache memory and a volatilization cache memory can also be used. In addition, in drawing 10 - drawing 12, a sign common to drawing 1 and a common component is attached, and explanation is omitted.

[0067] That is, in the case of drawing 10, two or more channel I/F control circuits 1a-1d by the side of a host and two or more disk control circuits 7a-7d by the side of a disk drive 8 consider

as the composition connected through the non-volatile cache memory 12 and the volatilization cache memory 13 which are arranged separately. Also in such composition, there is an advantage that the control circuit of the volatilization cache memory 13 can be simplified more with the effect in the composition illustrated by two or more channel I/F control circuits 1a-1d by [0068] In the composition to which between two or more disk control circuits 7a-7f by the side of a disk drive 8 was connected through the data path switch 11 in the case of drawing 11 The group of the independent non-volatile cache memory 12 and the volatilization cache memory 13 is mutually arranged for every sequence of two or more disk control circuits 7a, 7b, 7c, 7d, and 7e which make a sequence every disk drive 8, and 7f*8. In the composition of this drawing 11, there is an advantage that optimization of the capacity in the combination of the non-volatile cache memory 12 and the volatilization cache memory 13 or the combination of a throughput is realizable, for every sequence of a disk drive 8.

[0069] While omitting the data path switch 11 in drawing 11, in the case of drawing 12, each sequence of a disk drive 8 constitutes the parity group in the so-called RAID, and distributes and arranges the group of the independent non-volatile cache memory 12 and the volatilization cache memory 13 mutually for every parity group in it. In this case, when operation situations differ, for example for every parity group, there is an advantage that optimization of the capacity in the combination of the non-volatile cache memory 12 for every parity group concerned and the volatilization cache memory 13 or the combination of a throughput is realizable.

[0070] Although invention made by this invention person above was concretely explained based on the gestalt of operation, it cannot be overemphasized by this invention that it can change variously in the range which is not limited to the gestalt of the aforementioned implementation and does not deviate from the summary.

[0071] For example, it is widely applicable to the common data-storage system which has not only a disk subsystem but a memory hierarchy as a data-storage system.

[0072] [Effect of the Invention] According to the configuration method of the cache memory of this invention, the effect that the throughput of the data transfer which went via the cache memory can be raised is acquired.

[0073] Moreover, according to the configuration method of the cache memory of this invention, the effect that the cost reduction in the cache memory of composition of that a non-volatile cache memory and a volatilization cache memory are intermingled is realizable is acquired.

[0074] Moreover, according to the configuration method of the cache memory of composition of the effect that improvement in the mounting efficiency in the cache memory of composition of that a non-volatile cache memory and a volatilization cache memory are intermingled is realizable is acquired.

[Translation done.]

* NOTICES *

Japan Patent Office is not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1] It is the conceptual diagram showing an example of the composition of the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 2] It is the conceptual diagram having shown an example of the data flow at the time of execution of the lead command in the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 3] It is the conceptual diagram having shown an example of the data flow at the time of execution of the light command in the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 4] It is the conceptual diagram showing an example of the data flow at the time of two or more read/write generating which can be set to the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 5] In the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced, it is the conceptual diagram showing an example of data flow when two or more read/write concentrates on a specific disk drive.

[Drawing 6] It is explanatory drawing showing an example of control information used in the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 7] It is the flow chart which shows an example of the data write-in processing in the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 8] It is the flow chart which shows an example of the data write-in processing in the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 9] It is the flow chart which shows an example of the data read-out processing in the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 10] It is the conceptual diagram showing the modification of the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 11] It is the conceptual diagram showing the modification of the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 12] It is the conceptual diagram showing the modification of the data-storage subsystem by which the configuration method of the cache memory of this invention is enforced.

[Drawing 13] It is the conceptual diagram showing an example of the composition of a disk subsystem which consists of a disk controller of the former considered, and a subordinate's disk drive unit.

[Description of Notations]

1 (1a-1d) — A channel I/F control circuit, 2 (2a-2h) — Path, 3 [— Path,] (3a-3h) — A path, 4 (4a-4d) — Channel I/F, 5 (5a-5h) 6 [— Disk drive,] (6a-6h) — A path, 7 (7a-7h) — A disk control circuit, 8 (8a-8h) 9 — A disk control unit (memory control unit), 10 — Disk drive unit control circuit, 11 (11a, 11b) — A data path switch, 12 — A non-volatile cache memory, 13 [— (storage), 14 (14a, 14b) — A data path switch, 15 — A volatile cache memory, 16 [— Data flow, 17 — Data flow, 18-21 / — Data flow, 19-22 / — Data flow, 23-26 / — A NVS management flag 27 / — CM management flag,] (13a-13d) — 16 A volatilization cache memory, 17 — Data flow, 21-23

[Translation done.]

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平11-203201

(43)公開日 平成11年(1999)7月30日

(51)Int.Cl.⁶
G 0 6 F 12/08

識別記号
3 2 0

F I
G 0 6 F 12/08

G
3 2 0

審査請求 未請求 請求項の数3 OL (全13頁)

(21)出願番号 特願平10-2400

(22)出願日 平成10年(1998)1月8日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 森 健治

神奈川県小田原市国府津2880番地 株式会社日立製作所ストレージシステム事業部内

(74)代理人 弁理士 简井 大和

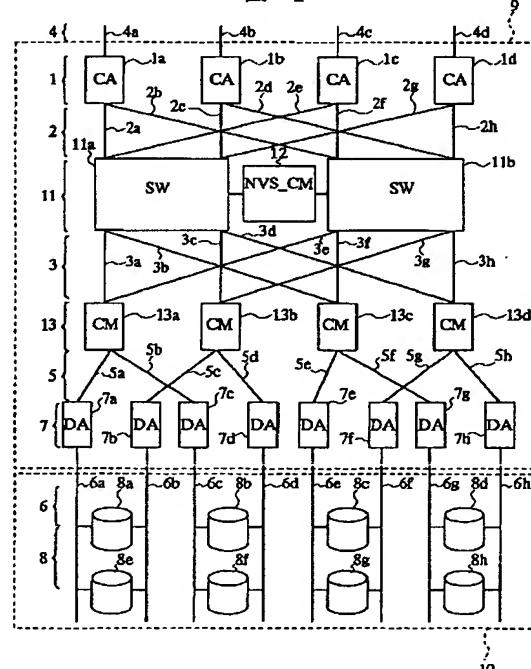
(54)【発明の名称】 キャッシュメモリの配置方法およびデータ記憶システム

(57)【要約】

【課題】 上位装置側のバス数や記憶装置数の増大に起因するキャッシュメモリへのアクセスのボトルネックを解消する。

【解決手段】 上位装置側の複数のチャネルI/F 4を制御する複数のチャネルI/F制御回路1と、ディスクドライブユニット10における複数のディスクドライブ8を制御する複数のディスク制御回路7との間に、上位装置側からの書き込みデータが格納される不揮発キャッシュメモリ12と、ディスクドライブ8から読出されたデータが一時的に格納される揮発キャッシュメモリ13を別個に設置するとともに、不揮発キャッシュメモリ12は、複数のデータ転送経路に共通に集中して設置し、揮発キャッシュメモリ13は、いくつかのデータ転送経路毎に分散して配置し、不揮発キャッシュメモリ12および複数の揮発キャッシュメモリ13の各々の容量やスループットを個別に設定可能にした。

図 1



【特許請求の範囲】

【請求項1】 上位装置と、前記上位装置との間で授受される情報が格納される記憶装置との間に配置され、前記情報を一時的に保持するキャッシュメモリの配置方法であって、

前記キャッシュメモリを不揮発キャッシュメモリおよび揮発キャッシュメモリにて構成し、前記不揮発キャッシュメモリは集中的に配置し、前記揮発キャッシュメモリは分散して配置することを特徴とするキャッシュメモリの配置方法。

【請求項2】 上位装置と、前記上位装置との間で授受される情報が格納される記憶装置との間に配置され、前記情報を一時的に保持するキャッシュメモリの配置方法であって、

前記キャッシュメモリを不揮発キャッシュメモリおよび揮発キャッシュメモリにて構成し、

前記不揮発キャッシュメモリの容量と、前記揮発キャッシュメモリの容量とが異なるように設定する第1の方法、

前記不揮発キャッシュメモリのスループットと、前記揮発キャッシュメモリのスループットとが異なるように設定する第2の方法、

の少なくとも一方の方法を用いることを特徴とするキャッシュメモリの配置方法。

【請求項3】 上位装置との間で授受される情報が格納される記憶装置と、前記記憶装置と前記上位装置との間における前記情報の授受を制御する記憶制御装置と、前記上位装置と前記記憶装置との間に配置され、前記情報が一時的に格納されるキャッシュメモリとを含むデータ記憶システムであって、

前記キャッシュメモリは、不揮発キャッシュメモリおよび揮発キャッシュメモリからなり、

前記不揮発キャッシュメモリは集中的に配置され、前記揮発キャッシュメモリは分散して配置される第1の構成、

前記不揮発キャッシュメモリの容量と、前記揮発キャッシュメモリの容量とが異なる第2の構成、

前記不揮発キャッシュメモリのスループットと、前記揮発キャッシュメモリのスループットとが異なる第3の構成、

の少なくとも一つの構成を備えたことを特徴とするデータ記憶システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、キャッシュメモリの配置技術およびデータ記憶技術に関し、特に、上位装置と記憶装置との間に設けられた複数の情報転送経路にて並列的に情報の授受が行われるデータ記憶システム等に適用して有効な技術に関する。

【0002】

【従来の技術】 図13は、考えられる従来のディスク制御装置109および配下のディスクドライブユニット110からなるディスクサブシステムの構成の一例を示す概念図である。従来のディスク制御装置109では、ホスト側の入出力制御を行うチャネル制御回路101と、ディスクドライブ側の入出力制御を行うディスク制御回路107との間に、キャッシュメモリアクセスバス105およびキャッシュメモリアクセスバス106を介してアクセスされるキャッシュメモリを配置する構成が一般的である。このとき、キャッシュメモリは、電源障害に起因するデータ喪失に備えた不揮発キャッシュメモリ102、またキャッシュメモリの障害時の補償用としての揮発キャッシュメモリ103、を持ち単純な2重系として使用していた。この場合、装置の対ホストに対するバス104や、ディスクドライブ108の数が増加した場合、全データアクセスにおいて2つの不揮発キャッシュメモリ102および揮発キャッシュメモリ103にアクセスが発生し、キャッシュメモリにアクセス集中するため、ディスク制御装置109の性能におけるボトルネックとなる。

【0003】 しかし従来のディスク制御装置109では対ホストに対するバス104、ディスクドライブ108の数が図13の様に4チャネル、8ドライブバス程度と少なく、この様な方式である程度の性能を確保できるため、この様な方式が取られてきた。

【0004】

【発明が解決しようとする課題】 今後ディスク制御装置当たりの容量の増加、高機能化により、制御を行うディスクドライブの数が多くなり、また装置のホストに対するバス数が増加すると考えられる。

【0005】 この様な制御装置を従来の方法で構成した場合、キャッシュメモリへのアクセスが集中し、キャッシュメモリが制御装置のボトルネックとなる。このボトルネックを解消するためには、キャッシュメモリを分散配置する方法が考えられる。この時、従来の様にそれぞれ障害に備え不揮発キャッシュメモリを含む2重系のキャッシュメモリを用意する必要があり、この様なもので分散キャッシュを構成すると実装的に困難で、また価格的にも高価となる。

【0006】 本発明の目的は、キャッシュメモリを経由したデータ転送のスループットを向上させることができ可能なキャッシュメモリの配置技術およびデータ記憶技術を提供することにある。

【0007】 本発明の他の目的は、不揮発キャッシュメモリと揮発キャッシュメモリとが混在する構成のキャッシュメモリにおけるコスト削減を実現することが可能なキャッシュメモリの配置技術およびデータ記憶技術を提供することにある。

【0008】 本発明の他の目的は、不揮発キャッシュメモリと揮発キャッシュメモリとが混在する構成のキャッシュ

シュメモリにおける実装効率の向上を実現することが可能なキャッシュメモリの配置技術およびデータ記憶技術を提供することにある。

【0009】

【課題を解決するための手段】本発明では、ライト時のデータを保証するための不揮発キャッシュメモリを1箇所に配置する。またリードの効率を上げるために、複数のアクセスができる様に揮発キャッシュメモリを分散配置する。この分散配置される揮発キャッシュメモリはリード時のみに使用し、データが万一電源切断等で揮発してもディスクからリードできるため、揮発性のメモリ媒体で構成可能である。

【0010】また、ディスク制御装置等の記憶制御装置ではリード/ライトの比率は対称でなく、リード4に対しライト1程度で使用される場合が多い。この性質を利用するとメモリ容量はリード/ライトの比率により不揮発キャッシュメモリはリード用の揮発キャッシュメモリほど大きくしなくとも、性能に対して影響しない。またスループットも不揮発キャッシュメモリはリード用の揮発キャッシュメモリより小さくできる。しかし分散配置するリード用の揮発キャッシュメモリの容量は敏感に性能に対して影響するため、不揮発キャッシュメモリよりも容量を大きくする必要がある。この様に、キャッシュメモリの配置を分散と集中の2つの組で、それぞれの容量を要求される性能に応じた値に設定することでキャッシュメモリを最適量にすることことができ、またこれにより実装的、価格的にも有利に、低価格で高性能化を図ることができる。

【0011】今後、ディスク制御装置当たりの容量の増加、高機能化により、制御を行うディスクドライブの数が多くなり、また装置のホストに対するバスが増加すると考えられる。

【0012】その場合、本発明では、複数のバスに対して不揮発キャッシュメモリを1箇所に集中的に配置し、揮発キャッシュメモリを、たとえば、一つあるいは幾つかのバス毎に分散配置する。これら揮発キャッシュメモリを分散することにより、同時にキャッシュメモリ転送が可能となり、多チャネル、多ドライブ化への対応時に1つのキャッシュメモリに集中することなく高いスループットをもたらすことができる。また分散するキャッシュメモリは揮発性でよいため実装的にも有利で、また低価格で構成可能である。さらに増設単位を分散配置される揮発キャッシュメモリのサイズを単位とすることにより、自由なスループット構成を探ることが可能となる。この様なキャッシュメモリの構成とすることで、たとえば複数バス構成のシステムにおいて低価格で高性能化を図ることが可能となる。

【0013】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照しながら詳細に説明する。

【0014】図1は本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムの構成の一例を示す概念図である。本実施の形態では、データ記憶サブシステムの一例として、ディスクサブシステムに適用した場合を例に採って説明する。

【0015】図1に例示される本実施の形態のディスクサブシステムは大きく、ディスク制御ユニット9、ディスクドライブユニット10、の2つのユニットより構成されている。ディスク制御ユニット9は、複数のチャネルI/F制御回路1(1a～1d)、複数のデータバススイッチ11(11a～11b)、不揮発キャッシュメモリ12、複数の揮発キャッシュメモリ13(13a～13d)、複数のディスク制御回路7(7a～7h)、より構成され、ディスクドライブユニット10は複数のディスクドライブ8(8a～8h)より構成されている。

【0016】図1において、複数のチャネルI/F制御回路1a～1dの各々は、複数のチャネルI/F4(4a～4d)を介して図示しないチャネル装置や中央処理装置(CPU)等のホスト装置に個別に接続され、チャネルI/F4のプロトコル制御、データ変換、データ転送を行う。

【0017】複数のデータバススイッチ11は、複数のバス2(2a～2h)および複数のバス3(3a～3h)を介して接続されている複数のチャネルI/F制御回路1と、不揮発キャッシュメモリ12および揮発キャッシュメモリ13との間におけるデータの流れの経路を切り替える役目をはたす。

【0018】複数の揮発キャッシュメモリ13は、複数のバス5(5a～5h)を介して複数のディスク制御回路7に接続され、複数のディスク制御回路7は、複数のバス6(6a～6h)を介してディスクドライブユニット10に接続されている。

【0019】不揮発キャッシュメモリ12は、ホスト側からのライト時のデータを格納するために用いる。これはライト処理の高速化の目的で、ホスト側からのライトコマンドに対して、不揮発キャッシュメモリ12への書き込みが完了した時点で、ディスクドライブ8に書き込む前に、完了を返す制御を行うため、万一、ディスクドライブ8にデータを書き込む前に停電で装置が停止した際のデータを、この不揮発キャッシュメモリ12でバックアップする目的に使用する。また、この不揮発キャッシュメモリ12に書き込む時は、可能な限り、同時に揮発キャッシュメモリ13に書き込みを行うようにデータバススイッチ11で経路の設定を行う。

【0020】揮発キャッシュメモリ13は主にリードデータのキャッシュメモリとして働き、ディスクドライブ8からデータをリードした際に、この揮発キャッシュメモリ13にデータを格納しておき、2回目以降に同じデータに対するリードに対してはディスクドライブ8から

データをリードするのではなく、揮発キャッシュメモリ13からリードすることにより高速化を図る目的に用いる。このためこの揮発キャッシュメモリ13は停電等により記憶データが揮発した場合でもディスクドライブにデータが保存されているため、揮発性のメモリ媒体で構成する事が可能である。

【0021】ディスク制御回路7は、バス6を介してディスクドライブ8に接続されており、ディスクドライブ8の制御を行うのに用いる。1つのディスク制御回路7は2つのディスクドライブ制御用のバスを有し、これにより万一の障害発生時にも交代バスを使用することによりディスクドライブ8へのアクセスが可能である。また通常時は2バスを有効に使用することで高速化の機能もはたす。

【0022】ディスクドライブ8は、データを格納するのに使用され、ディスクドライブ8へのアクセスバスは高速化、高信頼化のための2つの経路を有する。

【0023】すなわち、本実施の形態の場合には、ディスクドライブユニット10を構成する複数のディスクドライブ8は、いくつかのグループ（本実施の形態では一例として4つ）に系列化され、各系列は、複数のバス6a～6b、バス6c～6d、バス6e～6f、バス6g～6h、をそれぞれ介して、多重経路でディスク制御ユニット9に接続されている。また、複数のバス6a～6hの各々におけるデータ転送は、複数のディスク制御回路7a～7hの各々にて独立に制御される。

【0024】以下、本実施の形態のキャッシュメモリの配置方法およびディスクサブシステムの作用の一例について説明する。

【0025】図2はCPUからのリードコマンドの実行時のデータフローの一例を示した概念図である。この図2では、たとえばチャネルI/F制御回路1aにデータリードコマンドがチャネルI/F4aを介してCPUから入った場合について説明する。

【0026】まず、チャネルI/F制御回路1aでCPUからのコマンドを認識する。この結果リードコマンドであることがわかる。その後、どのディスクドライブ8のデータかを判別し、判別結果に応じて、たとえば、データバススイッチ11aにおいて切り替え、チャネルI/F制御回路1aと揮発キャッシュメモリ13aとを結合状態とする。

【0027】次にデータがキャッシュメモリ上にあるかの判定を行う。この判定はチャネルI/F制御回路1aで行う。この結果、対象データがキャッシュメモリ上にある（キャッシュヒット）と判定した場合には、データフロー16の様に揮発キャッシュメモリ13aよりチャネル制御回路1aにデータをデータバススイッチ11aを通じて転送することで読み出し、チャネルI/F制御回路1aがチャネルI/F4aを介してホスト（CPU）にデータを転送し、リードコマンドが終了する。

【0028】もしアクセス対象データが揮発キャッシュメモリ13a上に無い（キャッシュミス）場合は、データフロー17の様にディスク制御回路7a経由でディスクドライブ8aのリードを行い、データをデータバススイッチ11aを通じて転送し読み出す。この際、同時にデータを揮発キャッシュメモリ13aにライトする。これにより2回目以降のデータリードに対してはキャッシュヒットとなりアクセスの高速化を図ることができる。

【0029】図3はCPUからのライトコマンドの実行時のデータフローの一例を示した概念図である。チャネルI/F制御回路1aにチャネルI/F4aを介してデータライトコマンドがCPUから入った場合について説明する。

【0030】まずチャネルI/F制御回路1aでCPUからのコマンドを認識する。この結果ライトコマンドであることがわかる。この後ライト対象がどのディスクドライブ8かを判別し、判別結果に応じてデータバススイッチ11aを切り替え、たとえば、チャネルI/F制御回路1aと不揮発キャッシュメモリ12および揮発キャッシュメモリ13aを結合状態とする。

【0031】チャネルI/F制御回路1aからのデータをデータフロー21の様に同時に揮発キャッシュメモリ13a、不揮発キャッシュメモリ12に転送する。これによりライトデータは不揮発キャッシュメモリ12上にライトされるとともに、分散キャッシュメモリ13aに転送される。この後にバス5aを介してディスク制御回路7aにもライトデータが転送され、さらにバス6aを介して目的のディスクドライブ8aに転送されることによって当該ディスクドライブ8aにデータが書き込まれる。

【0032】次に、停電が発生した場合の動作について説明する。チャネルI/F制御回路1aがデータを受け取り、データフロー22の様に不揮発キャッシュメモリ12に転送が完了した時点でチャネルI/F制御回路1aはCPUに対しコマンド完了を返す。このためCPUではデータ書き込みが完了したと判断することになる。しかしこの時点ではディスク制御ユニット9内にデータが在り、まだディスクドライブ8aに書き込まれていない状態にある。このとき停電が発生すると、転送が完了した不揮発キャッシュメモリ12と揮発キャッシュメモリ13a上の内、揮発キャッシュメモリ13a上のデータは消えてしまう。この後、電源が復旧した時に、未書き込みのデータが不揮発キャッシュメモリ12上に存在するかの確認を行い、ディスク制御ユニット9内のデータのディスクドライブ8aへの書き込みが再開される。この時、データフロー23の様に不揮発キャッシュメモリ12からのデータは、データバススイッチ11aから、バス3a、揮発キャッシュメモリ13a、バス5a、ディスク制御回路7a、バス6a、を通じて目的のディスクドライブ8aへ転送される。

【0033】図4は複数のリード／ライト発生時のデータフローの一例を示す概念図である。

【0034】これはチャネルI／F制御回路1aにはチャネルI／F4aを介してライトコマンドの要求が、他のチャネルI／F制御回路1b、1c、1dにはチャネルI／F4b、4c、4dを介してリードコマンドの要求があった場合の処理の一例を示す図である。

【0035】まず、チャネルI／F制御回路1aの動作について説明する。チャネルI／F制御回路1aはライトコマンドを実行するため、データバススイッチ11aを制御して不揮発キャッシュメモリ12と揮発キャッシュメモリ13aへのバスを選択する。そしてCPUからのライトデータはデータフロー34の様に、バス2a、データバススイッチ11aを経由して不揮発キャッシュメモリ12に転送され、同時に、バス2a、データバススイッチ11a、バス3a、揮発キャッシュメモリ13a、バス5aを介してディスク制御回路7aに転送され、さらにバス6aを経由してディスクドライブ8aに書き込まれる。

【0036】次にチャネルI／F制御回路1b、1cの動作について説明する。チャネルI／F制御回路1bでリードコマンドを実行するため、データバススイッチ11aは揮発キャッシュメモリ13bを選択する。この時、データバススイッチ11aはチャネルI／F制御回路1aとチャネルI／F制御回路1bの2つの要求に応じ同時に2種類のバスを形成できる。そしてリード対象のデータが揮発キャッシュメモリ13bに存在した場合（キャッシュヒット）、データフロー35の様に、揮発キャッシュメモリ13bからバス3c、バス2cを経由してチャネルI／F制御回路1bへデータ転送が行われる。

【0037】また、チャネルI／F制御回路1cではリードコマンドを実行するため、データバススイッチ11bは揮発キャッシュメモリ13cを選択する。そしてリード対象のデータが揮発キャッシュメモリ13cに存在した場合（キャッシュヒット）、データフロー36の様に、揮発キャッシュメモリ13cから、バス3f、バス2fを経由してチャネルI／F制御回路1cへデータ転送が行われる。

【0038】最後にチャネルI／F制御回路1dの動作について説明する。チャネルI／F制御回路1dでリードコマンドを実行するため、データバススイッチ11bは揮発キャッシュメモリ13dを選択する。この揮発キャッシュメモリ13dに対象データが存在しない場合（キャッシュミス）、データフロー37の様にバス5hを経由してディスク制御回路7hにリード要求が入り、バス6hを経由してディスクドライブ8dから読出されたりードデータは揮発キャッシュメモリ13dにライトされながら、バス3h、バス2hを経由してチャネルI／F制御回路1dに転送される。この様に揮発キャッシュ

メモリ13を、バス毎に13a～13dのように分散配置することにより、複数のバスに対するデータ転送を同時にを行うことが可能となる。

【0039】図5は複数のリード／ライトが特定のディスクドライブ（たとえばディスクドライブ8d）に集中した場合のデータフローの一例を示す概念図である。

【0040】これはチャネルI／F制御回路1a、1bにライトコマンド、チャネルI／F制御回路1c、1dにリードコマンドの要求があった場合の図である。

【0041】まずチャネルI／F制御回路1aの動作について説明する。データフロー43の様に、チャネルI／F制御回路1aはライトコマンドを実行するため、データバススイッチ11aを不揮発キャッシュメモリ12と揮発キャッシュメモリ13cへのバスを選択する。しかしこの時、リードコマンド等によりディスク制御回路7gが使用されている場合は、不揮発キャッシュメモリ12までの転送のみを先に行い、その後、ディスク制御回路7gが空いてから、停電時と同じ様に不揮発キャッシュメモリ12から、揮発キャッシュメモリ13cを経由してディスク制御回路7gにデータを転送し、ディスクドライブ8dに書き込む。図5の例の場合は、ディスク制御回路7gが空いているため同時に転送を行う。

【0042】ホストに対するコマンドの終了報告はディスクドライブ8dに書き込みが完了する前に、不揮発キャッシュメモリ12に書込んだ時点で行う。基本的には不揮発キャッシュメモリ12へは確実に書き込み、その時、ディスク制御回路7gへのバスが空いている場合は同時にディスク制御回路7gへの転送を行う処理とする。

【0043】チャネルI／F制御回路1bの動作についても同様に、データフロー44の様に一時的に不揮発キャッシュメモリ12に書き込みを行い、その後、ディスク制御回路7hが空いてから不揮発キャッシュメモリ12から、揮発キャッシュメモリ13dを経由してディスク制御回路7hにデータを転送し、ディスクドライブ8dに書き込む。この例では不揮発キャッシュメモリ12に対する書き込みは、2つのチャネルI／F制御回路1a、1bよりのライト要求が競合するがこれは順番に行う。

【0044】次にチャネルI／F制御回路1cの動作について説明する。チャネルI／F制御回路1cでリードコマンドを実行するため、データバススイッチ11bが揮発キャッシュメモリ13dを選択する。そしてリード対象のデータが揮発キャッシュメモリ13dに存在した場合（キャッシュヒット）、データフロー45の様に揮発キャッシュメモリ13dから、バス3h、バス2fを経由してチャネルI／F制御回路1cへデータ転送が行われる。

【0045】もし揮発キャッシュメモリ13dに目的のデータが存在しない場合（キャッシュミス）は、バス6

h、ディスク制御回路7 h、バス5 h、揮発キャッシュメモリ13 d、バス3 hを経由してディスクドライブ8 dよりデータを読み出す。

【0046】チャネルI/F制御回路1 dについても同様に、データフロー46の様に、まず揮発キャッシュメモリ13 dにリード対象のデータがあるかの確認を行い、ある場合（キャッシュヒット）には揮発キャッシュメモリ13 dより転送を行い、無い場合（キャッシュミス）はディスクドライブ8 dよりデータを読み出す。この時、リードコマンドが競合するが、これは順番に行う。

【0047】従来の技術では、この様な場合、単一のディスクドライブに対するリード、ライトでも分散したディスクドライブに対するアクセスでも、キャッシュメモリに対するアクセスの競合が発生し、待ちが多くなることになり、性能低下の原因となった。

【0048】しかし本実施の形態では、分散したディスクドライブに対するアクセスでは競合は発生せず高性能が得られる。また図5の様に1つのディスクドライブ8 dに対するアクセスが集中するケースは、短いアクセスが多数発生するトランザクション処理では少ないため、リード時の特定の揮発キャッシュメモリに対するアクセスの集中は少ないといえる。また集中管理するメモリとして不揮発キャッシュメモリ12がありライト時にアクセスが集中するが、ライトの比率が一般的にリードに比べ4分の1程度のため、それほどの問題にはならない。そのため、これにより不揮発キャッシュメモリ12の容量も他の揮発キャッシュメモリ13 a～13 dの総和に比べて少なくすみ、低価格で高性能なディスクサブシステムを実現することが可能となる。

【0049】ここでキャッシュメモリのサイズ、スループットに関する説明を行う。図5の様に不揮発キャッシュメモリ12を1つ、また揮発キャッシュメモリ13を、一例として13 a、13 b、13 c、13 dの4組み持った構成の場合を考える。ライトとリードに対する比率を1:3とする。これによりに不揮発キャッシュメモリ12のサイズ、スループットはリードに用いる複数の揮発キャッシュメモリ13（13 a～13 d）の等価サイズ、スループットの4分の1程度ですむことになり、実装的、価格的に都合がよい。

【0050】次に複数個の揮発キャッシュメモリ13 a～13 dから構成される揮発キャッシュメモリ13のサイズ、スループットについて考える。この時、複数の揮発キャッシュメモリ13 a～13 dの各々のカーバー範囲がディスクドライブバス（バス5 aと5 b、バス5 cと5 d、バス5 eと5 f、バス5 gと5 h）とした場合、このディスクドライブバスに対する集中の度合いでその値が決まる。図5の様にバス5 a～5 hの8組で構成され、それぞれの競合がないとする場合、等価スループットそのもので良いが、平均2つの競合が発生すると

した場合は等価スループットの2倍の性能が必要になってくる。またサイズに関しては、単純には組み数（バス5の数）で割れば良いが、データの分散の程度により値が異なる。一部に集中した場合を想定した場合にはその分多く持つことでキャッシュメモリの効果を引き出すことができる。これは確率的に値を決めることになる。

【0051】ところで、本実施の形態のように、不揮発キャッシュメモリ12と、揮発キャッシュメモリ13（13 a～13 d）とを分散して配置し、不揮発キャッシュメモリ12に対するデータ書き込みが完了した時点でホスト側に書き込み完了を応答する構成では、不揮発キャッシュメモリ12への書き込みデータが、必ずしも直ちに揮発キャッシュメモリ13やディスクドライブ8に反映されているとは限らない。このため、未反映の間にリード要求が発生した場合には、最新のデータが、不揮発キャッシュメモリ12、揮発キャッシュメモリ13、ディスクドライブ8のいずれに存在するかを判別する操作が必要となる。

【0052】本実施の形態では、一例として、図6に例示されるような制御情報を用いて、このような判別操作を行う。

【0053】すなわち、不揮発キャッシュメモリ12では、たとえばアクセス単位のエントリ毎に、NVS管理フラグ50（V_N）を設ける。本実施の形態の場合、V_Nが“0”的とき、当該エントリの書き込みデータは揮発キャッシュメモリ13に未反映であり、“1”的ときは反映済である。

【0054】また、揮発キャッシュメモリ13では、たとえばアクセス単位のエントリ毎に、CM管理フラグ51（V, A）を設ける。本実施の形態の場合、Vが“0”的とき、当該エントリのデータに対して、不揮発キャッシュメモリ12に未反映のデータが存在し、Vが“1”的ときは存在しない。また、Aが“0”的とき、当該エントリの書き込みデータはディスクドライブ8上に未反映の状態にあり、Aが“1”的ときは反映済である。

【0055】なお、CM管理フラグ51においては、電源投入直後は、格納データが消失しているが、この状態では、全エントリのVおよびAは、ともに“0”的状態にあり、この状態では、キャッシュミスと判定され、ディスクドライブ8上からのデータリードが実行される。そして、不揮発キャッシュメモリ12に存在するディスクドライブ8に未反映のデータの当該ディスクドライブ8への書き込み操作や、ディスクドライブ8から読出されたデータの格納操作によって、VおよびAは後述のように変化する。

【0056】そして、データ書き込みに際しては、たとえば、図7のフローチャートに例示されるように、ホスト側（チャネルI/F制御回路1）から到来する書き込みデータを、不揮発キャッシュメモリ12に書き込んだ

のち、NVS管理フラグ50のV_Nを“0”にセットし（ステップ201）、さらにCM管理フラグ51のVを“0”にセットする（ステップ202）。その後、ホスト側にライト完了を応答する（ステップ203）。なお、ステップ202ではCM管理フラグ51のVの操作のために揮発キャッシュメモリ13へのアクセスが発生するが通常のデータ転送とは異なり、わずかなフラグビットの操作のみであるため、オーバーヘッドは少ない。

【0057】たとえば、上述の図5の例のように、不揮発キャッシュメモリ12以下への書き込みデータの転送は、任意契機でよく、たとえば、図8のフローチャート例示されるような手順にて行われる。

【0058】すなわち、まず、NVS管理フラグ50のV_Nが“0”的エントリを不揮発キャッシュメモリ12から検索し（ステップ301）、当該データを、揮発キャッシュメモリ13に転送した後、CM管理フラグ51のVを“1”にセットする（ステップ302）。さらに、揮発キャッシュメモリ13からディスクドライブ8上に書き込みデータを転送した後、CM管理フラグ51のAを“1”にセットする（ステップ303）。最後に、NVS管理フラグ50のV_Nを“1”にセットする（ステップ304）。この一連の操作は任意契機で実行可能である。

【0059】一方、任意の契機で発生するホスト側からのリード要求の処理は、一例として、図9に例示されるフローチャートのようにして行われる。

【0060】すなわち、リード要求が発生すると、まず該当する揮発キャッシュメモリ13のCM管理フラグ51がチェックされ（ステップ401）、A=1かつV=0の場合には、リード要求されたデータに対応した未反映の書き込みデータが不揮発キャッシュメモリ12に存在すると判定して、不揮発キャッシュメモリ12からデータを読み出してホストに転送する（ステップ404）。

【0061】また、ステップ401において、A=1かつV=0でないと判定された場合には、さらに、A=0かつV=1、または、A=1かつV=1か否かを調べ（ステップ402）、この条件が成立する場合には、揮発キャッシュメモリ13のキャッシュヒットとして、揮発キャッシュメモリ13内のデータを読み出してホスト側に転送する（ステップ405）。

【0062】ステップ401、ステップ402のいずれの条件にも合致しない場合には、キャッシュミスと判定し、ディスクドライブ8からデータを読み出し、揮発キャッシュメモリ13に書き込みつつ、ホスト側にデータを転送し、CM管理フラグ51のAおよびVを“1”にセットする（ステップ403）。

【0063】このようなNVS管理フラグ50およびCM管理フラグ51を用いた一連の処理により、データ書き込み要求に際してのデータ書き込み動作が、不揮発キャッシュメモリ12以下の揮発キャッシュメモリ13、

さらにディスクドライブ8のどのレベルで未実行であるか否かに関係なく、ホスト側からのリード要求に対し、最新データのリードを的確に実行可能であり、たとえば、最新の書き込みデータが未反映の古いデータを誤って読み出してホスト側に転送する、等の障害の発生を確実に回避することができる。

【0064】また、このような管理に際してアクセスされるデータは、高々数ビットであるため、NVS管理フラグ50およびCM管理フラグ51の操作に起因するオーバーヘッドはリード／ライト処理のスループットにはほとんど影響しない。

【0065】以上説明したように、本実施の形態のキャッシュメモリの配置方法およびデータ記憶システムによれば、複数のチャネルI/F4やバス2を備えた多チャネルバス化、ディスクドライブユニット10におけるディスクドライブ8の数量の増大によって多ディスクドライブ化されたディスク制御ユニット9の構成において、揮発キャッシュメモリ13をいくつかの経路毎に分散配置することで、高スループット化が可能となり、さらに、揮発キャッシュメモリ13とは別個に不揮発キャッシュメモリ12を集中的に配置して管理することで、不揮発キャッシュメモリ12および揮発キャッシュメモリ13の各々のサイズを最適に設定することができ、実装面で有利に、また低価格で高性能なディスク制御ユニット9、すなわち、ディスクサブシステムを実現することが可能となる。

【0066】なお、不揮発キャッシュメモリおよび揮発キャッシュメモリの分散配置方法としては、図1に例示された方法に限らず、たとえば、図10～図12に例示された構成を用いることもできる。なお、図10～図12において図1と共に構成要素には共通の符号を付して説明は割愛する。

【0067】すなわち、図10の場合には、ホスト側の複数のチャネルI/F制御回路1a～1dと、ディスクドライブ8側の複数のディスク制御回路7a～7dとが、別個に配置される不揮発キャッシュメモリ12および揮発キャッシュメモリ13を介して接続される構成としたものである。このような構成においても、上述の図1に例示される構成における効果とともに、揮発キャッシュメモリ13の制御回路をより簡略化できる、という利点がある。

【0068】図11の場合は、ホスト側の複数のチャネルI/F制御回路1a～1dと、ディスクドライブ8側の複数のディスク制御回路7a～7fとの間をデータバススイッチ11を介して接続した構成において、ディスクドライブ8毎に系列をなす、複数のディスク制御回路7a, 7b, 7c, 7d, 7e, 7f、の各系列毎に、互いに独立な不揮発キャッシュメモリ12および揮発キャッシュメモリ13の組を配置したものである。この図11の構成の場合には、ディスクドライブ8の系列毎

に、不揮発キャッシュメモリ12および揮発キャッシュメモリ13の組み合わせにおける容量やスループットの組み合わせの最適化を実現できる、という利点がある。

【0069】図12の場合には、図11におけるデータバススイッチ11を省略するとともに、ディスクドライブ8の各系列が、いわゆるRAIDにおけるパリティグループを構成し、各パリティグループ毎に、互いに独立な不揮発キャッシュメモリ12および揮発キャッシュメモリ13の組を分散して配置したものである。この場合には、たとえば各パリティグループ毎に稼働状況が異なる場合に、当該各パリティグループ毎の不揮発キャッシュメモリ12および揮発キャッシュメモリ13の組み合わせにおける容量やスループットの組み合わせの最適化を実現できる、という利点がある。

【0070】以上本発明者によってなされた発明を実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。

【0071】たとえば、データ記憶システムとしてはディスクサブシステムに限らず、記憶階層を有する一般的なデータ記憶システムに広く適用することができる。

【0072】

【発明の効果】本発明のキャッシュメモリの配置方法によれば、キャッシュメモリを経由したデータ転送のスループットを向上させることができる、という効果が得られる。

【0073】また、本発明のキャッシュメモリの配置方法によれば、不揮発キャッシュメモリと揮発キャッシュメモリとが混在する構成のキャッシュメモリにおけるコスト削減を実現することができる、という効果が得られる。

【0074】また、本発明のキャッシュメモリの配置方法によれば、不揮発キャッシュメモリと揮発キャッシュメモリとが混在する構成のキャッシュメモリにおける実装効率の向上を実現することができる、という効果が得られる。

【0075】また、本発明のデータ記憶システムによれば、キャッシュメモリを経由したデータ転送のスループットを向上させることができる、という効果が得られる。

【0076】また、本発明のデータ記憶システムによれば、不揮発キャッシュメモリと揮発キャッシュメモリとが混在する構成のキャッシュメモリにおけるコスト削減を実現することができる、という効果が得られる。

【0077】また、本発明のデータ記憶システムによれば、不揮発キャッシュメモリと揮発キャッシュメモリとが混在する構成のキャッシュメモリにおける実装効率の向上を実現することができる、という効果が得られる。

【図面の簡単な説明】

【図1】本発明のキャッシュメモリの配置方法が実施さ

れるデータ記憶サブシステムの構成の一例を示す概念図である。

【図2】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおけるリードコマンドの実行時のデータフローの一例を示した概念図である。

【図3】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおけるライトコマンドの実行時のデータフローの一例を示した概念図である。

【図4】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおける複数のリード/ライト発生時のデータフローの一例を示す概念図である。

【図5】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおいて、複数のリード/ライトが特定のディスクドライブに集中した場合のデータフローの一例を示す概念図である。

【図6】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおいて用いられる制御情報の一例を示す説明図である。

【図7】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおけるデータ書き込み処理の一例を示すフローチャートである。

【図8】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおけるデータ書き込み処理の一例を示すフローチャートである。

【図9】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムにおけるデータ読み出し処理の一例を示すフローチャートである。

【図10】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムの変形例を示す概念図である。

【図11】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムの変形例を示す概念図である。

【図12】本発明のキャッシュメモリの配置方法が実施されるデータ記憶サブシステムの変形例を示す概念図である。

【図13】考えられる従来の、ディスク制御装置および配下のディスクドライブユニットからなるディスクサブシステムの構成の一例を示す概念図である。

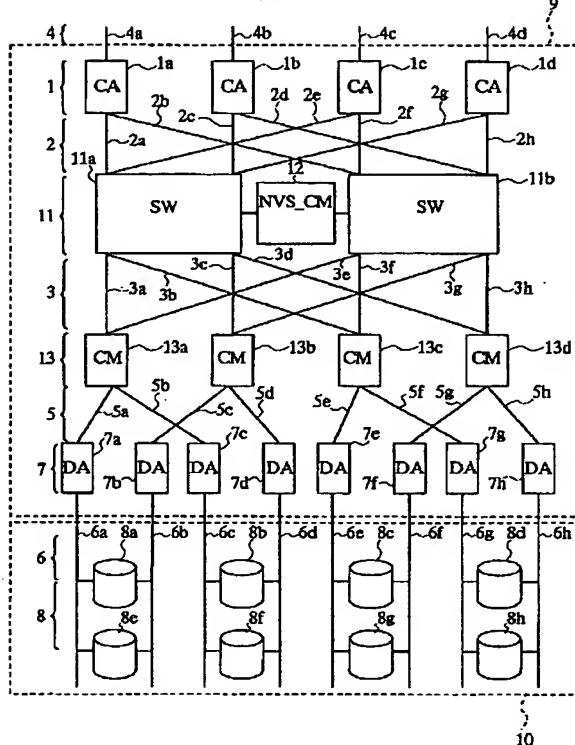
【符号の説明】

1 (1a～1d) …チャネルI/F制御回路、2 (2a～2h) …バス、3 (3a～3h) …バス、4 (4a～4d) …チャネルI/F、5 (5a～5h) …バス、6 (6a～6h) …バス、7 (7a～7h) …ディスク制御回路、8 (8a～8h) …ディスクドライブ、9…ディスク制御ユニット(記憶制御装置)、10…ディスクドライブユニット(記憶装置)、11 (11a, 11b) …データバススイッチ、12…不揮発キャッシュメモリ、13 (13a～13d) …揮発キャッシュメモリ、16, 17…データフロー、21～23…データフ

ロー、34～37…データフロー、43～46…データ
フロー、50…NVS管理フラグ、51…CM管理フラ
グ。

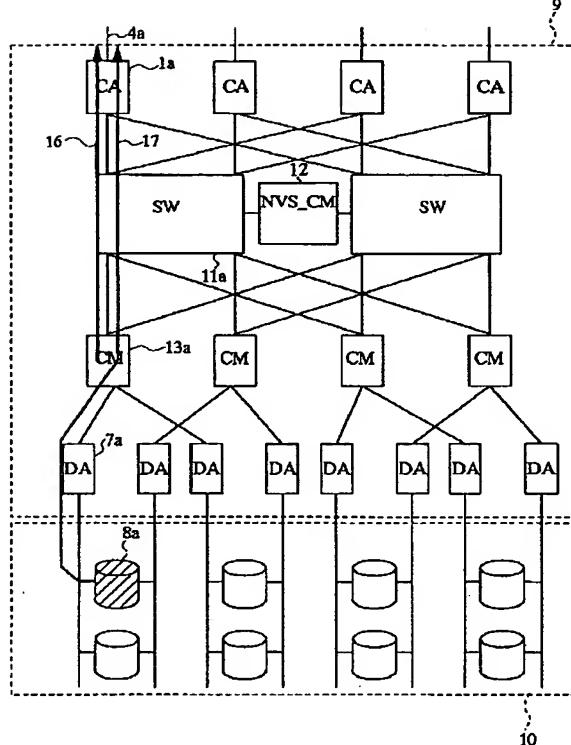
[図1]

图 1



〔図2〕

图 2



【図6】

6

NVS管理フラグ | V_N

$V_N = 0$: NVS上の書き込みデータはCMに未反映。
 $V_N = 1$: NVS上の書き込みデータはCMに反映済。

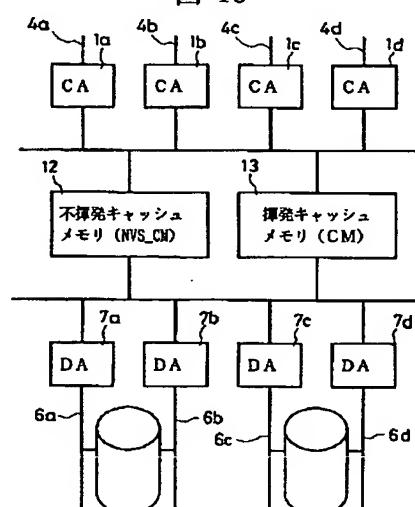
CM管理フラグ

V=0 : NVSからCMに未反映の書き込みデータあり。
V=1 : NVSからCMに未反映の書き込みデータなし。

A=0 : CMからD I S Kに未反映の書き込みデータあり。
A=1 : CMからD I S Kに未反映の書き込みデータなし。

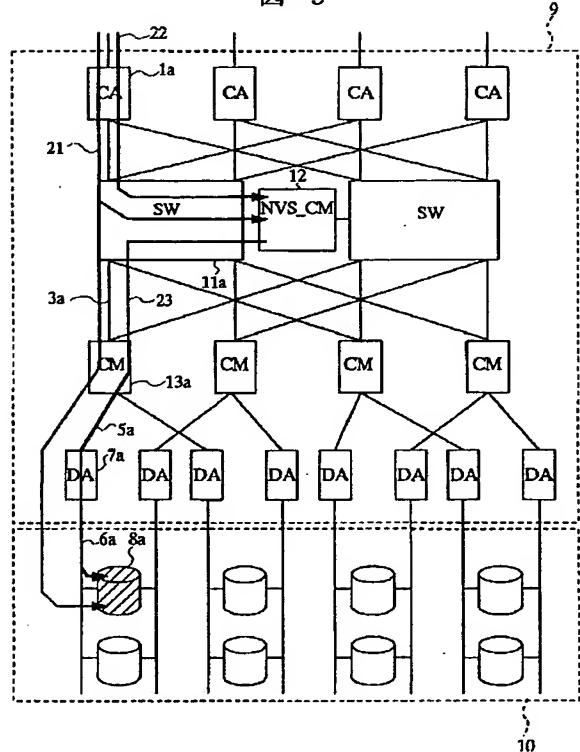
【図10】

図 10



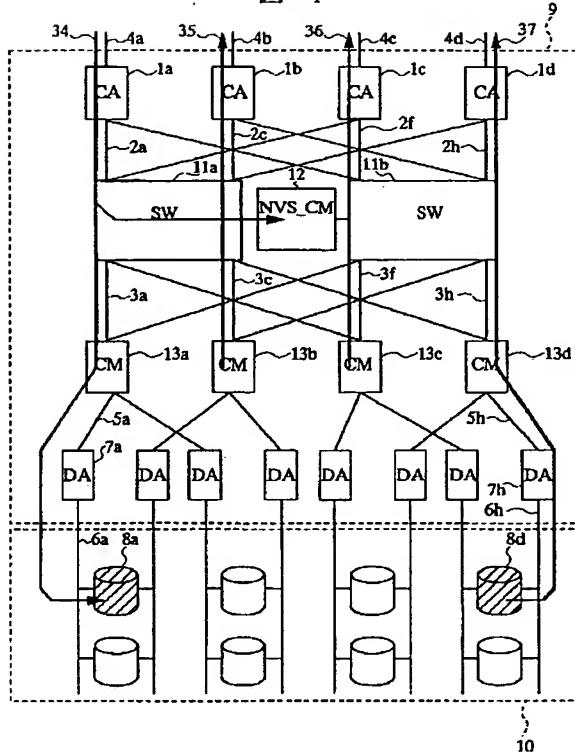
【図3】

図3



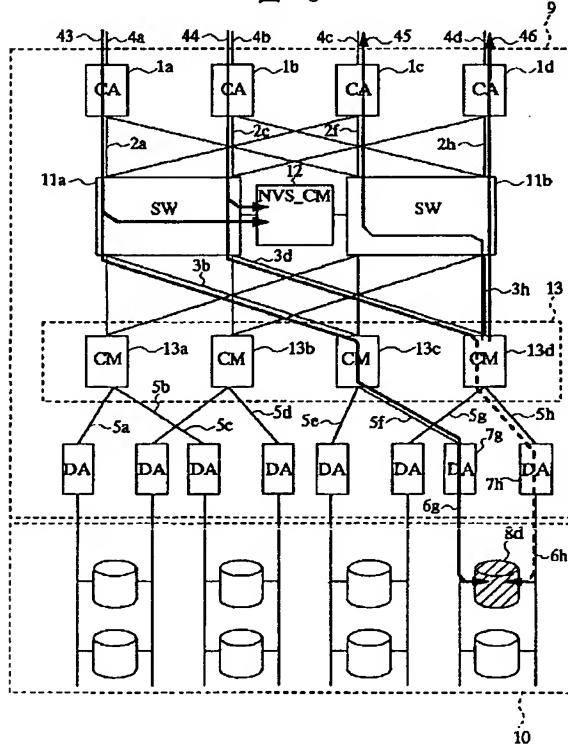
【図4】

図4

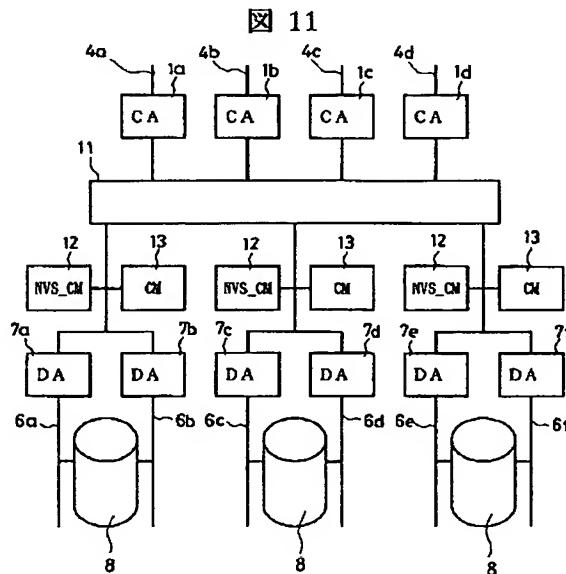


【図5】

図5



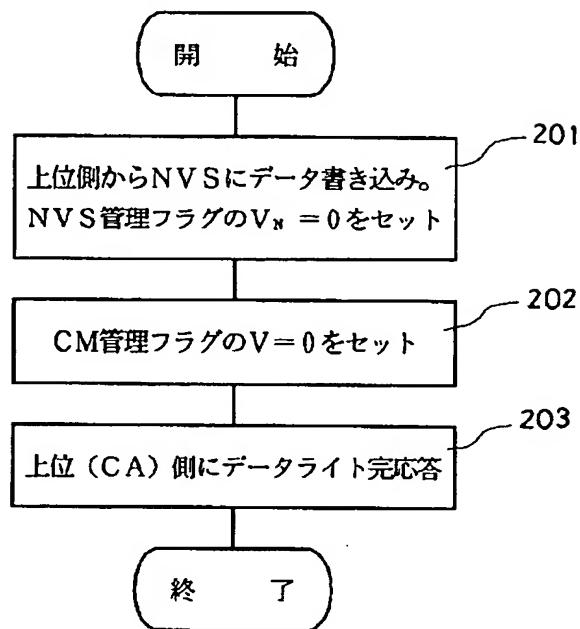
【図11】



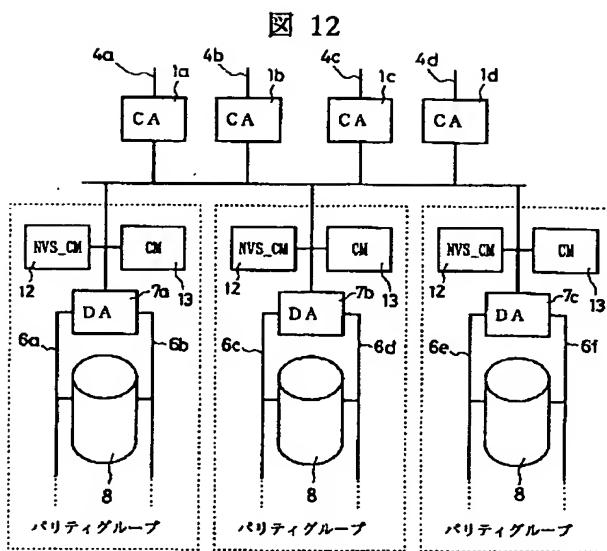
【図7】

図 7

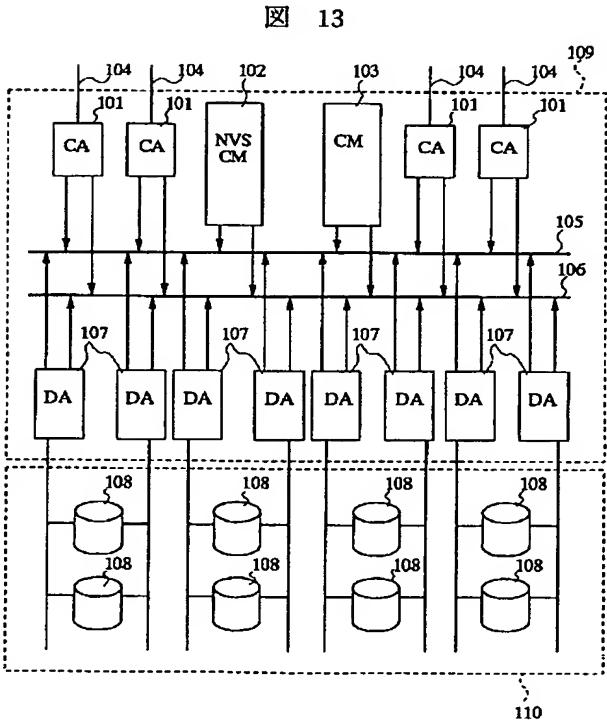
上位 (CA) 側からのデータ書き込み要求処理



【図12】



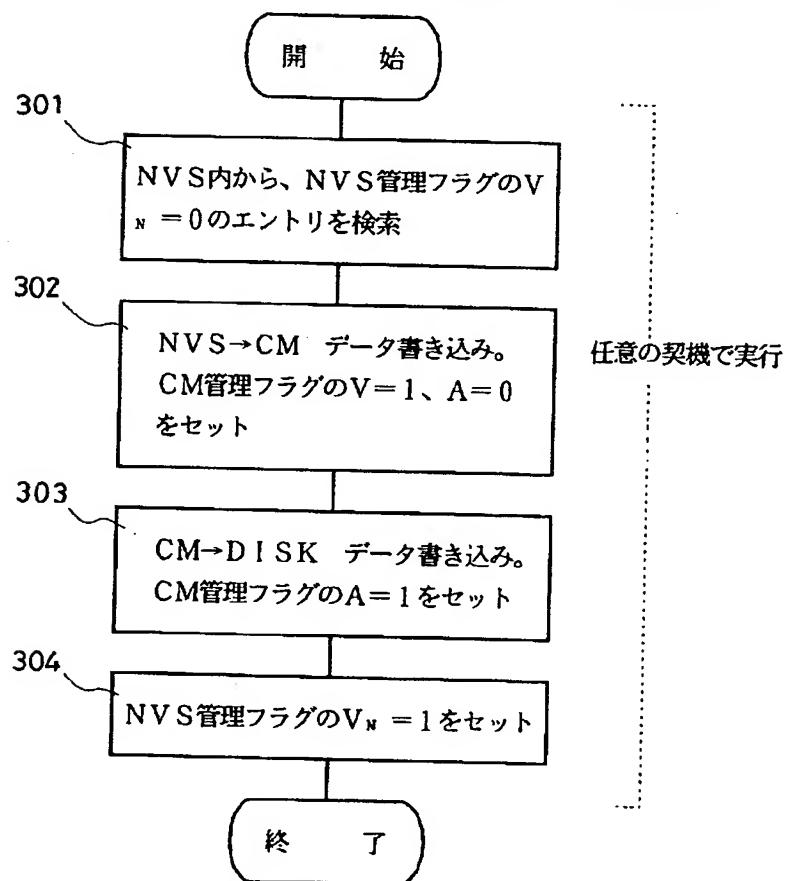
【図13】



【図8】

図 8

NVSからCMおよびDISK側への書き込みデータ反映処理



【図9】

図 9

上位(CA)側からのデータ読み出し要求処理

